A Tight Lower Bound on the Mutual Information of a Binary and an Arbitrary Finite Random Variable in Dependence of the Variational Distance

Arno G. Stefani, Johannes B. Huber
Institute for Information Transmission (LIT)
FAU Erlangen-Nuremberg
Erlangen, Germany
Email: {stefani, huber}@LNT.de

Christophe Jardin, Heinrich Sticht
Bioinformatics, Institute for Biochemistry
FAU Erlangen-Nuremberg
Erlangen, Germany
Email: {christophe.jardin, h.sticht}@biochem.uni-erlangen.de

***Abstract*—"THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD".**

In this paper a numerical method is presented, which finds a lower bound for the mutual information between a binary and an arbitrary finite random variable with joint distributions that have a variational distance not greater than a known value to a known joint distribution. This lower bound can be applied to mutual information estimation with confidence intervals.

I. Introduction

A tight lower bound for the mutual information between a binary and an arbitrary finite random variable with joint distributions that have a variational distance not greater than a known value to a known joint distribution can be found by minimizing over this set of joint distributions. Unfortunately, in general this minimization problem is hard to solve, since the mutual information is not convex in the joint distribution.

Therefore this minimization problem is split up into two subproblems.

If the marginal probability of the binary random variable is fixed, then the mutual information can easily be minimized over the conditional probabilities of the second random variable, since the mutual information is convex in the conditional probabilities [1, Theorem 2.7.4] and the set of conditional probabilities is convex (see Theorem 1) and therefore this optimization problem is convex. This constitutes the first subproblem which can easily be solved by standard methods for convex optimization.

In the second subproblem, having a closer look on the marginal probability distribution of the binary random variable, one first recognizes that this is only one-dimensional since the two probabilities have to sum up to 1. Next, the variational distance between the joint probabilities is greater or equal than the variational distance of the marginal probabilities, as is shown in (5). Therefore one can simply generate sufficiently many marginal probability distributions equidistantly in the one dimension left, solve the first subproblem for every of these marginal probability distributions and return the smallest mutual information calculated that way.

In the next section the notation is fixed. In section III the details of the method are given. In section V some numerical examples are shown.

II. Notational setup

Let X, Y be a pair of finite discrete random variables, with joint probability distribution

$$p_{XY} = \{p_{XY}(i,j) : i = 1, 2, \dots, M_x;\; j = 1, 2, \dots, M_y\}.$$

Here $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ and it is w.l.o.g. assumed that $\mathcal{X} = \{1, 2, \dots, M_x\}$ and that $\mathcal{Y} = \{1, 2, \dots, M_y\}$. The marginal probability distributions are $p_X = \{p_X(i) : i = 1, 2, \dots, M_x\}$ and $p_Y = \{p_Y(j) : j = 1, 2, \dots, M_y\}$. They are calculated from the joint probalility distributions as usual. The conditional probability distributions are

$$p_{Y|X} = \{p_{Y|X}(j|i) : i = 1, 2, \dots, M_x;\; j = 1, 2, \dots, M_y\},$$
$$p_{X|Y} = \{p_{X|Y}(i|j) : i = 1, 2, \dots, M_x;\; j = 1, 2, \dots, M_y\}.$$

It is defined that $p_{Y|X}p_X = p_X p_{Y|X} = p_{X|Y}p_Y = p_Y p_{X|Y} = p_{XY}$. The product of the marginal distributions is denoted as

$$p_X p_Y = \{p_X(i)p_Y(j) : i = 1, 2, \dots, M_x;\; j = 1, 2, \dots, M_y\}.$$

For any two joint probability distributions p_{XY}, q_{XY} the relative entropy or Kullback-Leibler distance [1] is defined as

$$D(p_{XY} \| q_{XY}) = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} p_{XY}(i,j) \log \frac{p_{XY}(i,j)}{q_{XY}(i,j)} \quad (1)$$

and the mutual information between X and Y [1] as the relative entropy between the joint probability distribution and product of the marginal probability distributions of X and Y

$$I(X;Y) = I(p_{XY}) = D(p_{XY} \| p_X p_Y). \quad (2)$$

All $\log$s are assumed to be natural if not stated otherwise.

The variational distance between two joint probability distributions is defined as

$$V(p_{XY}, q_{XY}) = \|p_{XY} - q_{XY}\|_1 = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} |p_{XY}(i,j) - q_{XY}(i,j)|,$$

and similarly for the marginal distributions. It can be easily seen, that $V(\cdot,\cdot) \in [0,2]$ for any two probability distributions.

III. Results

First it is shown that set of all conditional probability distributions constrained by a maximal variational distance is convex.

Theorem 1: Let $p_{XY} = p_X p_{Y|X}$ be any fixed joint probability distribution of any two two discrete finite random variables X, Y, let q_X be any fixed probability distribution of X and let ϵ be any fixed number $\in [0,2]$. Then the set $\mathcal{Q} = \{q_{Y|X} \mid V(q_X q_{Y|X}, p_{XY}) \le \epsilon\}$ is convex.

Proof: Let $q^1_{Y|X}, q^2_{Y|X}$ be any two conditional probability distributions $\in Q$. Then one only has to show that the convex combination $q^\lambda_{Y|X} = \lambda q^1_{Y|X} + (1-\lambda) q^2_{Y|X}$, with $\lambda \in [0,1]$ is also in Q. Before this is done, it is defined that $q^1_{XY} = q_X q^1_{Y|X}$, $q^2_{XY} = q_X q^2_{Y|X}$ and $q^\lambda_{XY} = \lambda q^1_{XY} + (1-\lambda) q^2_{XY} = q_X q^\lambda_{Y|X}$. Now, to proof that $q^\lambda_{Y|X} \in \mathcal{Q}$, one only has to show that $V(q_X q^\lambda_{Y|X}, p_{XY}) \le \epsilon$. Herefore

$$\begin{aligned} V(q_X q^\lambda_{Y|X}, p_{XY}) &= V(q^\lambda_{XY}, p_{XY}) \\ &= \left\| q^\lambda_{XY} - p_{XY} \right\|_1 \le \epsilon, \end{aligned} \tag{3}$$

where the fact that any norm ball is convex [2, Section 2.2.3] has been used in (3). Also, the further constraints implied by the probability simplex (which is convex) are no problem since an intersection of convex sets is always convex [2, Section 2.3.1]. ■

Since the empty set is convex, no restriction on $V(p_X, q_X)$ (e.g. $V(p_X, q_X) \le \epsilon$) is necessary.

Corollary 1: Let p_{XY} be any fixed joint probability distribution of any two two discrete finite random variables X, Y, let q_X be any fixed probability distribution of X and let ϵ be any fixed number $\in [0,2]$. Then, the optimization problem

$$\min_{q_{Y|X} \;:\; V(q_X q_{Y|X}, p_{XY}) \le \epsilon} I(q_X q_{Y|X}) \tag{4}$$

is convex.

Proof: The mutual information $I(q_X q_{Y|X})$ is a convex function of the conditional probabilities $q_{Y|X}$ when q_X is fixed, and the set $\{p_{Y|X} \mid V(q_X q_{Y|X}, p_{XY}) \le \epsilon\}$ is convex. ■

Corollary 1 basically says that the optimization problem given is practically solvable. However, since it is a general convex optimization problem, it can still be cumbersome to find a suitable algorithm with the correct parameters. Fortunately the problem can be restated in such a way, that it can be handled by disciplined convex programming (DCP) [3], which works perfectly well for this problem as can be seen in section V.

The minimization problem in Corollary 1 can not be solved in a straightforward manner with DCP, since this would violate the no product rule of DCP (see (1), (2)), also there is no built function in CVX (which is the software which implements DCP) for the mutual information as a function of the conditional probabilities when the corresponding marginal probability is fixed. Therefore the relative entropy, which is a built in function in CVX and is convex in its two input arguments, is used. Then it can be seen that

$$I(X;Y) = I(q_X q_{Y|X}) = D(q_X q_{Y|X} \| q_X q_Y),$$

and $q_X(i) q_{Y|X}(j|i)$ are affine functions of $q_{Y|X}(j|i)$ as $q_X(i) q_Y(j) = q_X(i)(\sum_i q_{Y|X}(j|i) q_X(i))$ are. Hence, the convexity of $D(\cdot,\cdot)$ is preserved [2, section 2.3.2], and it is straightforward to implement the minimization problem in Corollary 1 with CVX with this knowledge.

Next the second subproblem, namely the minimization of the mutual information over the marginal probability distribution q_X, is solved. Herefore it is first shown that

$$\begin{aligned} V(q_X, p_X) &= \|q_X - p_X\|_1 \\ &= \sum_{i=1}^{M_x} |q_X(i) - p_X(i)| \\ &= \sum_{i=1}^{M_x} \left| \sum_{j=1}^{M_y} (q_{XY}(i,j) - p_{XY}(i,j)) \right| \\ &\le \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} |q_{XY}(i,j) - p_{XY}(i,j)| \\ &= V(q_{XY}, p_{XY}) \\ &\le \epsilon. \end{aligned} \tag{5}$$

Therefore only q_X with $V(q_X, p_X) \le \epsilon$ have to be considered. Until here all results are applicable to any finite M_x, but from here the restriction $M_x = 2$ applies. In this case q_X is one dimensional obviously, and the set of all q_X is simply $\{q_X = \{\min(p_X(1) + \gamma, 1), \max(p_X(2) - \gamma, 0)\} \mid \gamma \in [-\frac{\epsilon}{2}, \frac{\epsilon}{2}]\}$. Practically, the minimization problem

$$\min_{q_{XY} \;:\; V(q_{XY}, p_{XY}) \le \epsilon} I(q_{XY}) \tag{6}$$

is then simply solved by generating sufficiently many q_X equidistantly in γ, solve the optimization problem of Corollary 1 for every q_X and return the smallest mutual information calculated that way. Here the number of q_Xs is considered to be sufficient if one gets a smooth graph for the mutual information minimized over the conditional probabilities $q_{Y|X}$ as a function of γ.

IV. Discussion

Together with the bound on the probability of a maximal variational distance between the true joint distribution and an empirical joint distribution (see [6], and especially an refinement of it which drops the dependence on the true distribution [4, Lemma 3]) the given bound can be used to construct a reasonably tight lower bound of the confidence interval for mutual information. Such an application can be found in [8]. In mutual information estimation with confidence intervals, the bound given is especially useful, when the marginal probability distribuition is far from being uniform. Such a situation can be found in [7]. In the case of two binary random variables the results seem to coincide with lower bound of [5].

V. NUMERICAL EXAMPLES

In the first example (Fig. 1) a distribution p_{XY} and a maximal variational distance ϵ was handpicked to show that the mutual information minimized over the transitional probabilies $q_{Y|X}$ as a function of γ is neither convex nor concave (even for two binary random variables) and seems to be not differentiable at $\gamma = 0$, as can be seen in Fig. 1. The parameters chosen therefore are

$$p_{XY}(1,1) = 0.017, p_{XY}(1,2) = 0.285$$
$$p_{XY}(2,1) = 0.424, p_{XY}(2,2) = 0.274$$
and $\epsilon = 0.3$.

Then,

$$I(p_{XY}) \approx 0.2210 \text{ and}$$
$$\min_{q_{XY}\,:\,V(q_{XY},p_{XY})\leq\epsilon} I(q_{XY}) \approx 0.0019.$$

In all figures I is equal to the minimum of $I(q_X q_{Y|X})$ over $q_{Y|X}$ for fixed $q_X = \{\min(p_X(1)+\gamma, 1), \max(p_X(2)-\gamma, 0)\}$, constrained by $V(q_X q_{Y|X}, p_{XY}) \leq \epsilon$, and 1000 points were generated equidistantly for $\gamma \in [-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$.

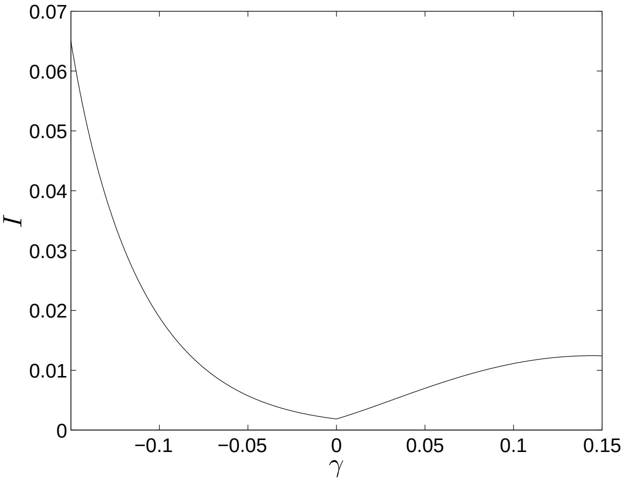

Fig. 1.

In the second example (Fig. 2) $M_y = 5$ and the following joint distribution was chosen at random (rounded for easier reproducibility)

$$p_{XY}(1,1) = 0.090, p_{XY}(1,2) = 0.098, p_{XY}(1,3) = 0.207,$$
$$p_{XY}(1,4) = 0.064, p_{XY}(1,5) = 0.026,$$
$$p_{XY}(2,1) = 0.239, p_{XY}(2,2) = 0.030, p_{XY}(2,3) = 0.104,$$
$$p_{XY}(2,4) = 0.107, p_{XY}(2,5) = 0.035,$$
and $\epsilon = 0.1$.

Then,

$$I(p_{XY}) \approx 0.1112 \text{ and}$$
$$\min_{q_{XY}\,:\,V(q_{XY},p_{XY})\leq\epsilon} I(q_{XY}) \approx 0.0524.$$

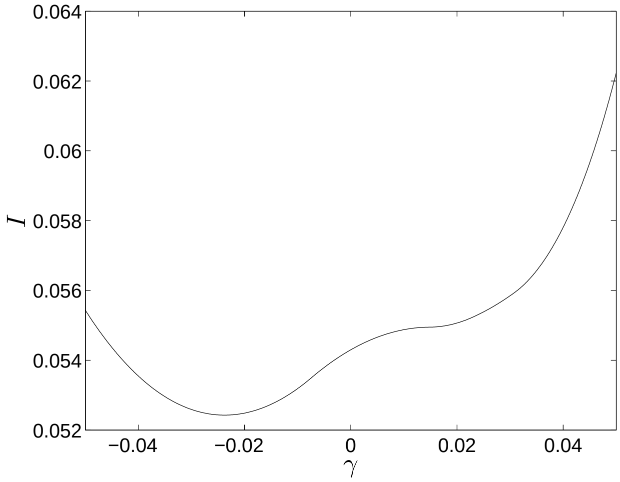

Fig. 2.

In the last example (Fig. 3) $M_y = 10$ and the following joint distribution was chosen at random (rounded for easier reproducibility)

$$p_{XY}(1,1) = 0.101, p_{XY}(1,2) = 0.062, p_{XY}(1,3) = 0.025,$$
$$p_{XY}(1,4) = 0.088, p_{XY}(1,5) = 0.005, p_{XY}(1,6) = 0.007,$$
$$p_{XY}(1,7) = 0.069, p_{XY}(1,8) = 0.059, p_{XY}(1,9) = 0.080,$$
$$p_{XY}(1,10) = 0.074,$$
$$p_{XY}(2,1) = 0.103, p_{XY}(2,2) = 0.006, p_{XY}(2,3) = 0.038,$$
$$p_{XY}(2,4) = 0.002, p_{XY}(2,5) = 0.018, p_{XY}(2,6) = 0.079,$$
$$p_{XY}(2,7) = 0.049, p_{XY}(2,8) = 0.032, p_{XY}(2,9) = 0.020,$$
$$p_{XY}(2,10) = 0.020,$$
and $\epsilon = 0.1$.

Then,

$$I(p_{XY}) \approx 0.1311 \text{ and}$$
$$\min_{q_{XY}\,:\,V(q_{XY},p_{XY})\leq\epsilon} I(q_{XY}) \approx 0.0369.$$

ACKNOWLEDGMENT

The authors would like to thank the DFG for supporting their research with SPP1395 in the projects HU634_7 and STI155_3.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[3] M. C. Grant, S. Boyd, and Y. Ye, *CVX: Matlab Software for Disciplined Convex Programming*, 2005. Available at http://www.stanford.edu/~boyd/cvx/
[4] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.

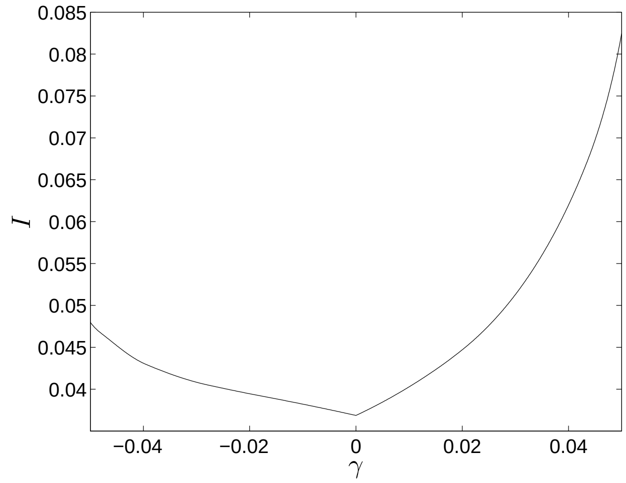

Fig. 3.

[5] A. G. Stefani, J. B. Huber, C. Jardin and H. Sticht, "Towards confidence intervals for the mutual information between two binary random variables," In *Proc. Workshop Computational Systems Biology (WCSB 2012)*, Ulm, Germany, Jun. 4–6, 2012.

[6] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M.J. Weiberger, "Inequalities for the L1 Deviation of the Empirical Distribution," Tech. Rept., HP Laboratories Palo Alto, HPL-2003-97 (R.1), Jun. 2003.

[7] O. G. Othersen, A. G. Stefani, J. B. Huber and H. Sticht, "Application of Information Theory to Feature Selection in Protein Docking," *J Mol Model.*, vol. 18, no. 4, pp. 1285–1297, Jul. 2012.

[8] A. G. Stefani, J. B. Huber, C. Jardin and H. Sticht, "Confidence intervals for the mutual Information," available at http://arxiv.org/abs/1301.5942.